A Grid based solution for Management and Analysis of Microarrays in distributed Bone Marrow Stem Cells experiments

Beltrame F (1), Corradi L (1), Milanesi L (2), Papadimitropoulos A (1), Porro I (1), Scaglione S (1), Schenone A (1), Torterolo L (1), Viti F (1)

(1) Department of Computer Science, Control Systems and Telecommunications- DIST-University of Genoa, Italy
(2) Biomedical Technologies Institute (ITB), National Research Council, Segrate, Milano, Italy

Motivation

Exploitation of gene expression data is fully dependent on the availability and sharing of genomic data and advanced statistical analysis tools, which are typically collected on distributed databases/providers and structured under different standards. For these reasons, a Grid based Environment for distributed Microarray data Management and Analysis (GEMMA) is presented. Different microarray (m.a.) analysis algorithms will be offered to the end-user through web interface. A set of independent applications will be published on the portal, and either single algorithms or a combination of them might be invoked by the user, through a workflow strategy. The services will be implemented within an existing grid computing infrastructure to solve problems concerning both the large datasets storage (data intensive problem) and the implied large computational time (computing intensive problem). Moreover, experimental data annotations will be collected according to the same criteria and stored through the Grid portal by using a metadata schema, allowing a comprehensible and replicable sharing of m.a. experiments available in GEMMA among different researchers.

Methods

As first stage a Grid portal will be released, based on Genius. Genius is nowadays a standard of graphical user interface access to the EGEE and Italian Grid infrastructures so it appears as the most suitable and convenient solution in our implementation. To complete this process, applications will be provided as grid services exploiting standard Grid infrastructure (authentication, inter-process communication, data management, job scheduling). From a functional point of view, the adopted framework allows to deploy both servlet components (visible to users as traditional web pages) and services (grid or web-services) exposing key components to the public with standard interfaces. From a data point of view, the proposed environment permits users to upload/download their data and results on/from the Grid Portal and store them on Grid storage resources. GEMMA environment is based on LCG, the official middleware for the Italian Grid infrastructure. A remarkable metadata management system is provided by the ARDA Metadata Catalogue Project (AMGA) metadata management server and clients that can be easily integrated in the LCG environment. The system is directly accessible through the web interface where users fill a multipage web form with required fields. The form is then processed and submitted to the catalog, as soon as the uploading of data on Grid provides a consistent logical reference to the data themselves. In order to develop a functional distributed management environment, particular attention has been addressed to the description of experiments using standard annotations. MIAME/MAGE guidelines have been adopted, omitting some standard fields and joining others to focus on our specific biological domain. DChip, one of the most complete and diffuse free software for the m.a. data analysis, was chosen for our application to cover tasks from image normalization to chromosome location, passing through filters to eliminate redundant or useless information, sample comparison extracting differentially expressed genes, hierarchical clustering or LDA classification to facilitate bioinformatics analysis. To enhance scalability in our implementation, the use of alternative open source sw for the genomic data analysis and comprehension, such as bioconductor, has also been considered. The proposed environment will be tested by using a specific experimental scenario: human bone marrow stem cells (BMSC) exposed to different experimental conditions.

Results

This platform provides a shared, standardized and reliable storage of biological data related to BMSC culture. Different m.a. analysis algorithms will be offered to the end-user, through a portal

web interface, starting from the Linux-ported dChip analysis tool. Several applications can be invoked and combined by the user, through a workflow strategy. Problems concerning large datasets storage, storage safety, and large computational times are solved through a grid computing infrastructure. The Grid portal will act as user interface for data storage, metadata management, data analysis and result retrieval. Data access from processing job will be done directly on distributed file system, without moving m.a. datasets to computing nodes local filesystem. Experimental data annotation will be gathered in GEMMA and stored with the use of a metadata scheme facilitating the apprehension and sharing of m.a. experiments. The tool can be also used to test several algorithms with different parameter configurations on the same dataset simultaneously and it's open to the integration of third-party modules for specific functions (like clustering algorithms and statistical analysis tools). This platform is currently part of the Italian FIRB project LITBIO (Laboratory for Interdisciplinary Technologies in BIOinformatics).

Contact email: pivan@bio.dist.unige.it